

Unsupervised Classification

“Clustering”

MultiSpec PC

A Beginning Note: The MultiSpec software is an ongoing project of the *Laboratory for Applications of Remote Sensing*, at Purdue University. Developed by David Landgrebe and programmed by Larry Biehl, the software was originally written for the Macintosh platform. The software is being converted to the Windows platform and, while almost complete, there are some features of the Macintosh version that are not yet implemented for PC machines. These will eventually be included and it is recommended that the user visit the MultiSpec home page (see page 5) from time-to-time for updates. This tutorial was written for the June 6, 1997 release of the software. Users of older versions of PC MultiSpec may find that the Clustering process is not “active,” or available. To complete this tutorial, visit the MultiSpec web site and download the newest version. The GLOBE web pages provide a link to this site.

To use this tutorial, you will need two image files; the bevsub.lan and bevsub.cls. These are available from the GLOBE Program website. They may also be downloaded from:

<http://ekman.unh.edu/globe/>

Select the folder “MultiSpec.” The files, bevsub.lan and bevsub.cls will work with either the Macintosh or PC versions of MultiSpec.

Unsupervised Classification

“Clustering” -- PC Version - Windows

Each pixel in your Landsat TM image contains a wealth of information about the surface materials that reflected light from that pixel to the satellite sensors. Each pixel contains a value which can range from 0 to 255, for each TM band supplied with your image. If, for instance, your image contains data for five bands, then each pixel contains five pieces of data, each potentially ranging from 0 to 255, as shown in the sample pixel diagram to the right.

This means that your image could contain 256^5 (that's approximately 1.1 billion) different possible spectral combinations. Each of these combinations does **not** represent a different type of land cover; most of these variations represent very small and, to us, “unseeable” differences in surface reflectance.

In most instances, your computer monitor will be displaying only 256 different colors, hence only 256 different pixels. Even set to “thousands” of colors, only a small part of the many different pixels can be displayed. Even if a monitor could display all the different possible pixels, your eyes could recognize only a small number of differences in their appearance.

Because there is a limited number of different land cover types (the Modified UNESCO Classifications scheme, MUC, contains about 157 different types), and no GLOBE study site will have all of those different land cover types, it is necessary to group pixels together into a smaller number of closely related “classes.” This process, whereby pixels with similar spectral characteristics are grouped, is called “Classification,” and is done in two different ways.

In a supervised classification, you “train” the software to recognize that certain types of pixels represent specific land cover types. This is done on the basis of your knowledge of your own area, and field work you may do. The software then classifies the pixels of your image into the groups you have specified.

In an unsupervised classification, or “Clustering”, we enter the number of groups, or “clusters,” we wish to have, and certain other specifications. The software then examines the pixels in the image and groups them according to similar spectral characteristics. These groupings are not made on the basis of land cover, but on the similarity of the spectral characteristics of the pixels.

As part of your preparation of a land cover map for your 15 km x 15 km GLOBE Study Site, it is necessary for you to identify relatively large, homogeneous areas in your image for ground study and later use in a supervised classification. To do this, you will have MultiSpec cluster your image. This will help you locate areas to visit for ground verification studies.

Landsat Pixel

Band 1	Blue	39
Band 2	Green	53
Band 3	Red	25
Band 4	Near IR	129
Band 5	Mid IR	46

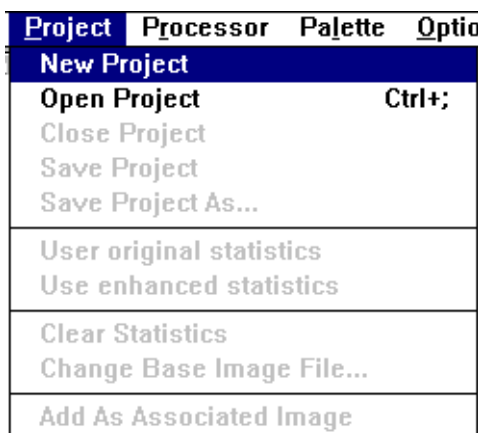
30 m

30 m

Clustering

To demonstrate clustering, you will use a “sub-set” of the Beverly, Massachusetts image provided with your MultiSpec tutorial. This 101 x 101 pixel sub-image will allow the demonstration process to proceed more quickly than the clustering of a 512 x 512 image, and will allow you to follow exactly the steps outlined in this tutorial.

- Launch **MultiSpec** and **Open** the **bevsub.lan** image. This is a sub set of a 5 channel, multispectral, Landsat Thematic Mapper image.
- From the **Project** menu, as shown below, select **New Project**.



Your clustering exercises are saved as “Projects” and, when done, can be opened by MultiSpec as “Thematic Images.”

- From the **Processor** menu, select **Cluster**. “Clustering” is MultiSpec’s terminology for an Unsupervised Classification. As shown on the next page, The **Set Cluster Specifications** window opens. It is in this window that you select a clustering “algorithm” (method by which the software clusters) and enter certain values for the software to use.

Set Cluster Specifications

Algorithm

☐ Single Pass...

☒ ISODATA...

Channel: All Available

Cluster Classification Map Area(s)

☐ No classificaiton map

☐ Training Area(s)

☒ Image Area

Area to Classify

	Start	End	Interval
Line	1	101	1
Column	1	111	1

Classification threshold 16

Write Cluster Report/Map To:

☒ Project Text Window

☒ Disk File

Symbols: Default set

Cluster Stats: Add To Project

Cancel OK

- First, click the **Image Area** button to place a dot in it.
- Click to place a marker in the **Disk File** box. This saves your project to disk.
- Lastly, click the **ISODATA** button, as indicated by the cursor in the diagram above. **ISODATA** is the algorithm, or mathematical process, that MultiSpec will use in the clustering process.

A new window, the **Set ISODATA Cluster Specifications** window will open, as shown below.

Set ISODATA Cluster Specifications

Initialization Options

- ☒ Along first cov. eigenvector
- ☐ Along first cor. eigenvector
- ☐ Within eigenvector volume
- ☐ Use single-pass clusters

Other options

Number clusters:

Convergence (%):

Minimum cluster size:

Determine clusters from

- ☐ Training Area(s)
- ☒ Image Area

Area to Cluster

Line:

Column:

Cancel

OK

It is in this window that you tell MultiSpec how you want the clustering to proceed. The information you need to provide is:

- Be certain that the **Image Area** radio button is checked, as shown above.
- Select “**Along first cov. eigenvector.**” This is the specific *algorithm*¹ that MultiSpec will use in its clustering
- Leave the settings in the **Other options** boxes unchanged for this exercise.

Notes: “Number of clusters” tells the software how many different groups you wish for the classification. The number 10 is used, for now, because we are clustering a small area. The number you will use when you cluster your 512 x 512 image will be discussed later.

During the classification, the program goes through the data over and over. This is called “iteration.” Each iteration is called a “pass”. The system makes “passes” through the image until a preset percentage of the pixels in the image are **not** changed during the pass. The clustering then ends. This percentage is called the “**Convergence.**”

“**Minimum cluster size**” tells the system the smallest sized area to work with. Areas smaller than this minimum size will not be clustered.

¹ For a discussion of MultiSpec’s algorithms, see “An Introduction to MultiSpec,” by David Landgrebe and Larry Biehl, Purdue Research Foundation, 1995. This document may be downloaded from the Purdue/LARS WWW site at:

- After you have made these settings, click **OK**.
- The **Set Cluster Specifications** window appears *again*.

Set Cluster Specifications

Algorithm

☐ Single Pass...

☒ ISODATA...

Channel: All Available

Cluster Classification Map Area(s)

☐ No classificaiton map

☐ Training Area(s)

☒ Image Area

Area to Classify

	Start	End	Interval
Line	1	101	1
Column	1	111	1

Cluster Stats: Add To Project

Write Cluster Report/Map To:

☒ Project Text Window

☒ Disk File

Classification threshold 100

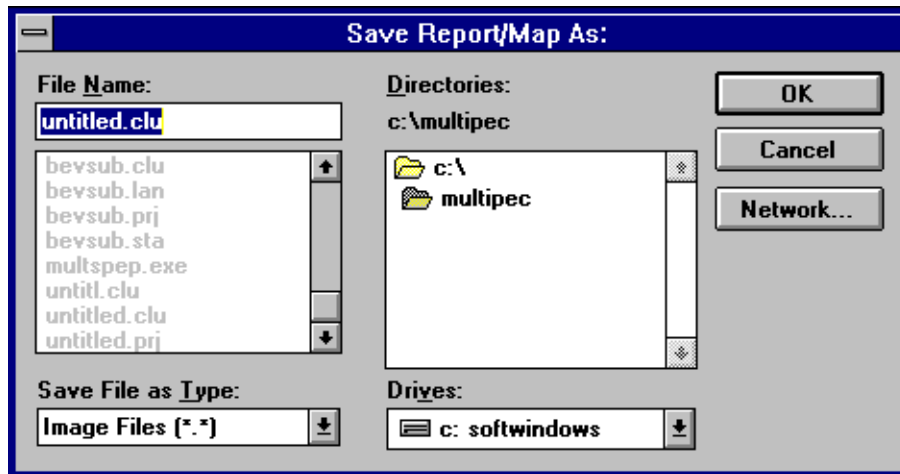
Cancel OK

- In the lower left-hand corner of the box is the “**Classification threshold:**” entry box. **Change the value in this box to “100”** just as you would change any item in a word-processor.

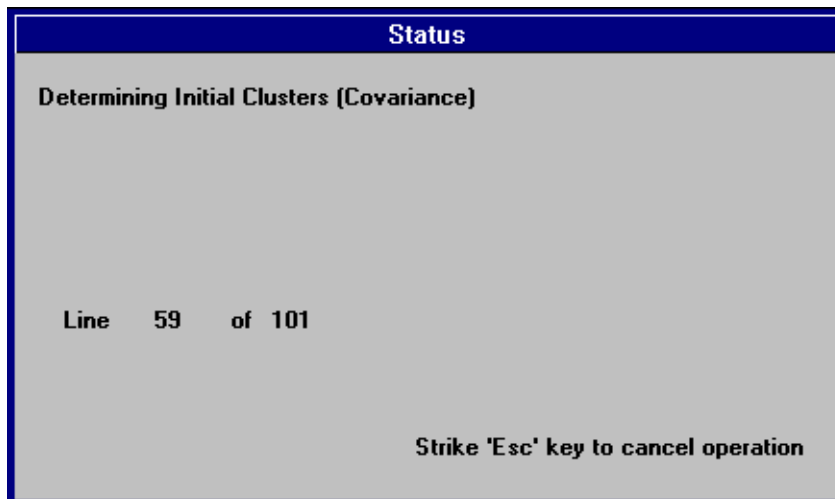
Setting this “threshold” value to 100 forces the system to assign every pixel in the image to one of the clusters. A value of less than 100 specifies the tolerance for assignment of pixels. A value of less than 100 will result in some pixels not being assigned to clusters. In this clustering, you are interested in large, fairly homogeneous areas, so individual pixels of slightly different spectral characteristics dotting the map are unnecessary.

- Click **OK**.

- The **Save Report/Map As:** dialog box appears, as shown below. There is a default name for your classified image file “**untitled.clu**.” You should change the “Untitled Project” portion to something more descriptive, but leave the “.clu” extension to tell you and the system what type of file this is.



- The system then makes its first pass through the image to initially determine the clusters present as shown in the **Status** box, below.



- The “Pass 1” clustering Status box then appears, as shown below. During this initial iteration, Pass 1, the “Percent of Pixels Not Changed” shows no value. Also note that a time is given for completion of this operation.

Status		
ISODATA Cluster - Pass 1.		
Line:	38	of 101
Percent of pixels not changed:		
Minutes until completion:	0.1	
Strike 'Esc' key to cancel operation		

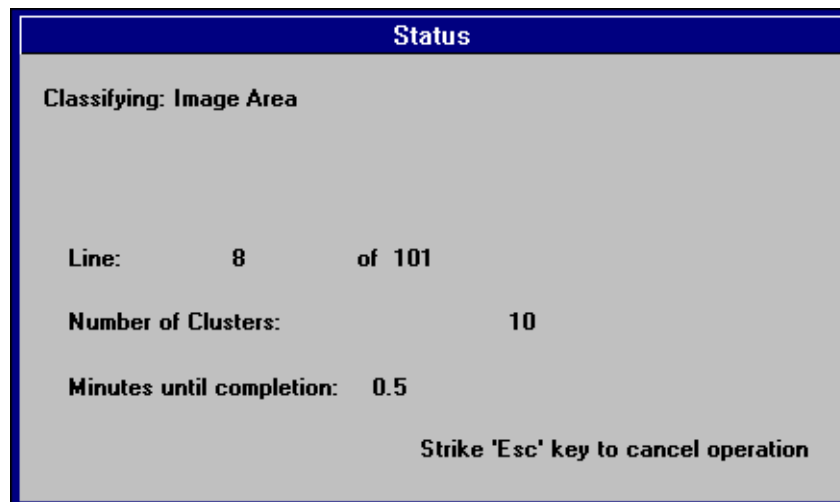
- The **Percent of Pixels Not Changed** entry does not change until the end of Pass 2. During Pass 3, the value will be displayed, as shown below, along with a time to completion of the pass.

Status		
ISODATA Cluster - Pass 3.		
Line:	29	of 101
Percent of pixels not changed:	54.2	
Minutes until completion:	0.1	
Strike 'Esc' key to cancel operation		

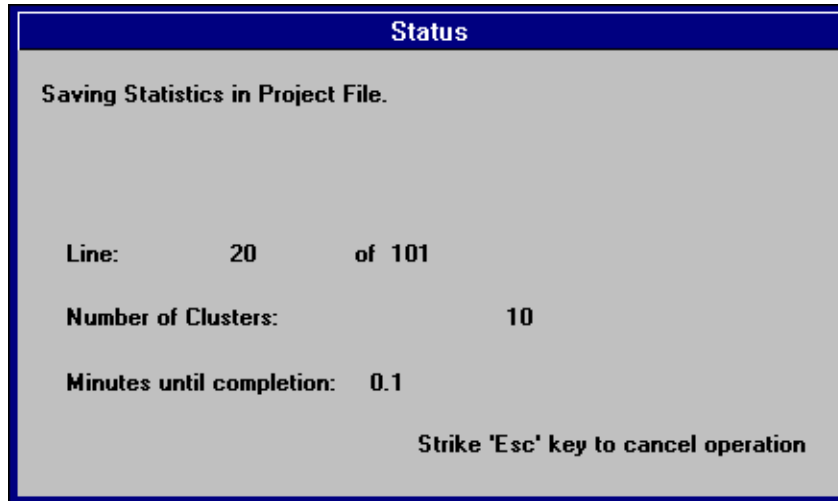
- During subsequent passes, the **“Percentage of Pixels Not Changed”** increases, until it reaches the value given in the **“Convergence (%)” specification**. The time for each pass to be completed is given in this window.

You can expect the system to make several passes to achieve a 98% Convergence. The time required for this process is dependent upon the processing speed of your computer. On a 386-based machine (the minimum processor requirement), you can expect the entire process to take about 5 minutes for this “sub” image. On a Pentium machine, the process is done very quickly. A full 512 pixel by 512 pixel GLOBE Study Site image will take longer to process.

- If you press **“Cancel”** during a pass, you will be asked if you wish to cancel immediately, or complete the iteration. Canceling immediately terminates the clustering, while finishing the iteration ends the clustering at a Convergence less than initially specified.
- After the clusters are determined, the system will display the **“Classifying Selected Image Area”** window, below. Here the system assigns individual image pixels to the clusters it has determined.



- After the clustering is complete, you will see the **Saving Statistics in Project File** window, shown below.



The Results of Clustering

There are two results of clustering:

A **description of clustering activity** and a “text map” in the **TEXT OUTPUT** window,

A **clustered Thematic image**.

- From the **Windows** menu, select **untitled project**. Scroll to the top of this “text window”, and you will have statistics describing the clustering and its results. A part of the text output for the sample clustering is shown below. In it are listed the number of clusters produced and the average value (mean) of the pixel values for each band in each of the classes.

```

Clustering completed after 8 passes and 35 of 2243 pixels changed.
Final cluster class statistics.
Cluster  Pixels  Channel Means
      1      2      3      4      5
1      267      94.7 101.4  85.6 178.7 152.3
2      129     123.0 149.3 125.2 228.0 238.0
3      194      73.3  91.3  63.2 240.4 172.1
4      429      63.5  78.5  52.3 218.7 148.7
5      493      61.0  70.5  48.0 191.2 130.8
6      290      62.4  66.3  48.4 150.1 108.1
7      183     129.3 126.4 115.9 138.9 138.6
8      118     182.0 171.6 164.1 130.4 152.3
9       66     243.8 243.0 243.8 140.3 201.8
10      74      62.6  46.7  26.9  17.0  13.5

```

Also produced is a text map of the clustered area. The system assigns a number or letter to each of the clusters, and then displays a map of the clustered area using this code. For the clustered **bevsub.lan** image, the code is shown below.

```

Number classes = 11

Classes used:
0: Thresholded
1: Cluster 1           1
2: Cluster 2           2
3: Cluster 3           3
4: Cluster 4           4
5: Cluster 5           5
6: Cluster 6           6
7: Cluster 7           7
8: Cluster 8           8
9: Cluster 9           9
10: Cluster 10        A

```

A portion of the Text Map from the Clustering process. Each number/letter represents a pixel and the clustered group to which it belongs. You can see, even from this representation, that the system has identified several large, homogeneous areas, identified by the appearance of the same letter/number in an area.

```

Classification of Selected Area
Lines 1 to 101 by 1. Columns 1 to 111 by 1
44445544444444444444555555555558765554444455171566665444445434555544434443775445533333
544455444444444343334444555555555551666544334444511345665545554433355544433332764444344434
5445544433333333334444555555555555266544433444345511335655555442997117133333337764443345555
4545544433444433444444555555555558764333433554441134455665548999879874433333765444444565
444554344444444344554344555555555511543334334454454355566655199999881444333327654344444554
454443454444334544334455555555551443334434444543455666655199999991443333371543334445444
544433454444344543334455445555555115444444444544555666555189999999133333175543455545433
444433355554444444333455555555555114555555455445556665544189888999143331765541155544333
4443335665554444433333466655556654477155115555444166665129999988998543337154441154444333
4443356665444444333336AA65555555561986651111111416666199999998998533332754411115444333
44334566654444433333466655555466648866771717811117766899889999974333176445515165444333
44333666555433333335655545665566433118999265767777718989999999991333715454441154444333
4333366554444333333455555566555653331789992651655318899899999999973375444431755443333
433334555444333333345555556655565333179992146655434799889999999984327144434445555443333
43333455554333333333355555665554554333119999136554334188189989922243371444441154515543333
4344445555433333333346555554544554345418991556544334511178722843333874444454314186444334
3344445543113333333455555554455455445447866655114334578821128143337154444451815554344444
4344444541713333333455555555545654455551775512881435511122281443281444554445655414455444
1444554555443233344555555665555433355545175517155446189998345432815444445445565198455455
144555554443333333333555556655566533444444417541881444899999544387144444554445651787114455
33455454443333333455555565456664334433344778778741999999131871145455653316A61788213415
344444333333344344455555545665433444443341871341711999999877111133445554116AA6112214445
1544454433333333333444445555556533444444344387433417199987776556133115545561AA6641117871
5545554443333333333455456655656545444333433714434122217787666556325121454551666555118881
55444334333333321555546665455544544433344417544417888776651655321522145551566466111115
5554411113333344455554456643444445544433344437111788776666651555323542225555564216666555
5444413122333344445433554333441544334333311287788771666665544555321653221555543216666655
1444333321333311777717777177777771277788888888711517665665434543221654222455511666555555
1433333277888888777777787777888888788788877181431156665555435543221555322165651554455445

```

- If you scroll to the bottom-right portion of this text window, you can see, as shown to the right, that the clustering process has used the letter “A” to represent the ocean area.

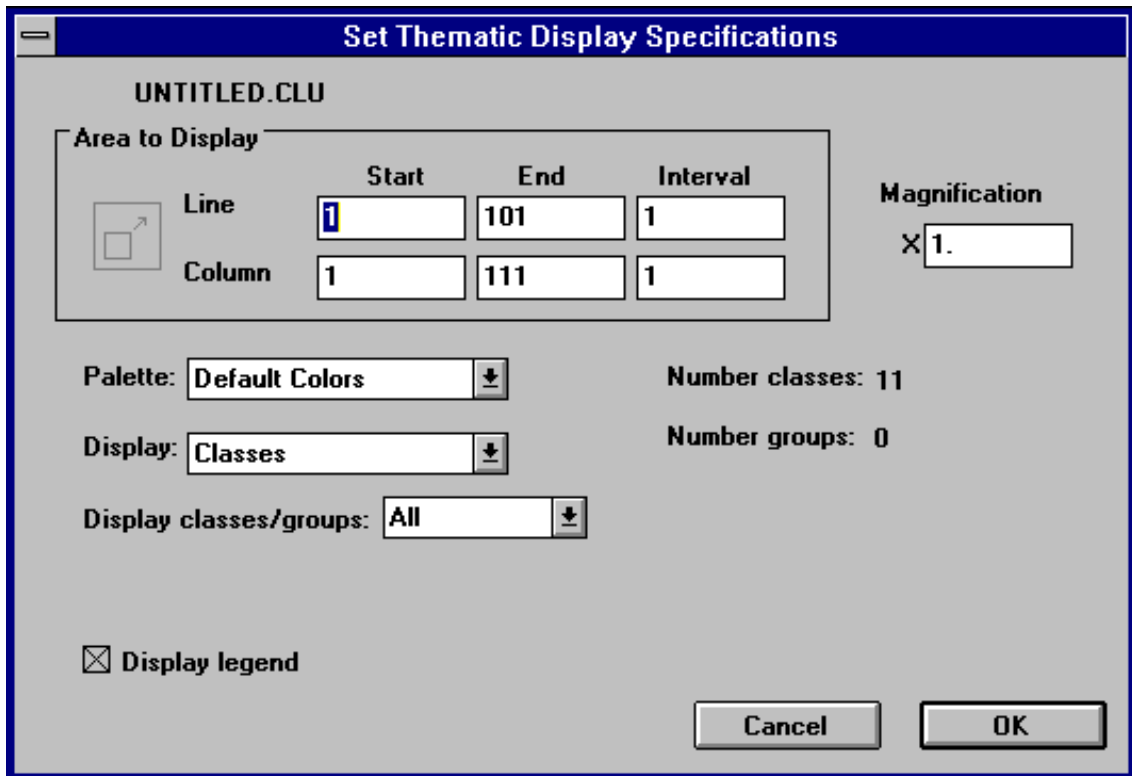
```

544555555655665554444566111111133322899
544565544554555444445665511189999999999
544566544455454444441165111799999999999
434556544455554444511111799999999888AA
444555444544565554551131999999988AAAAA
545666515445565565566611999998AAAAAAA
4566666715455655655554519998AAAAAAA
4456666776555656655544111298AAAAAAA
3455555176556665441155111297AAAAAAA
33565556445555544775417899AAAAAAA
515545555456551144665112997AAAAAAA
51511551545554115565418999AAAAAAA
543155545455443114551298AAAAAAA
13311555511441214518199AAAAAAA
121715434654411444899999999999999999999
122215411113311117998988888888888888888
3221155511128877189799999999999999999999
233466641129987779888888888888888888888
23345651129997A7997999999999999999999999
14445511299888887999999999999999999999999
54144129999999999999999999999999999999999
7897299999888888888888888888888888888888
9999999998888888888888888888888888888888


```

Examining the Clustered Image

- From the File menu, select **Open Image**.
- Select the **.clu** file name you used earlier, and click **Open**.
- The **Set Thematic Display Specifications** window opens, as shown below. You can experiment later with some of the other palettes in this menu, but for now accept the default settings and press **OK**.



The dialog box titled "Set Thematic Display Specifications" for the file "UNTITLED.CLU" contains the following settings:

Area to Display		Start	End	Interval
Line		1	101	1
Column		1	111	1

Magnification: X 1.

Palette: Default Colors (dropdown arrow)

Display: Classes (dropdown arrow)

Display classes/groups: All (dropdown arrow)

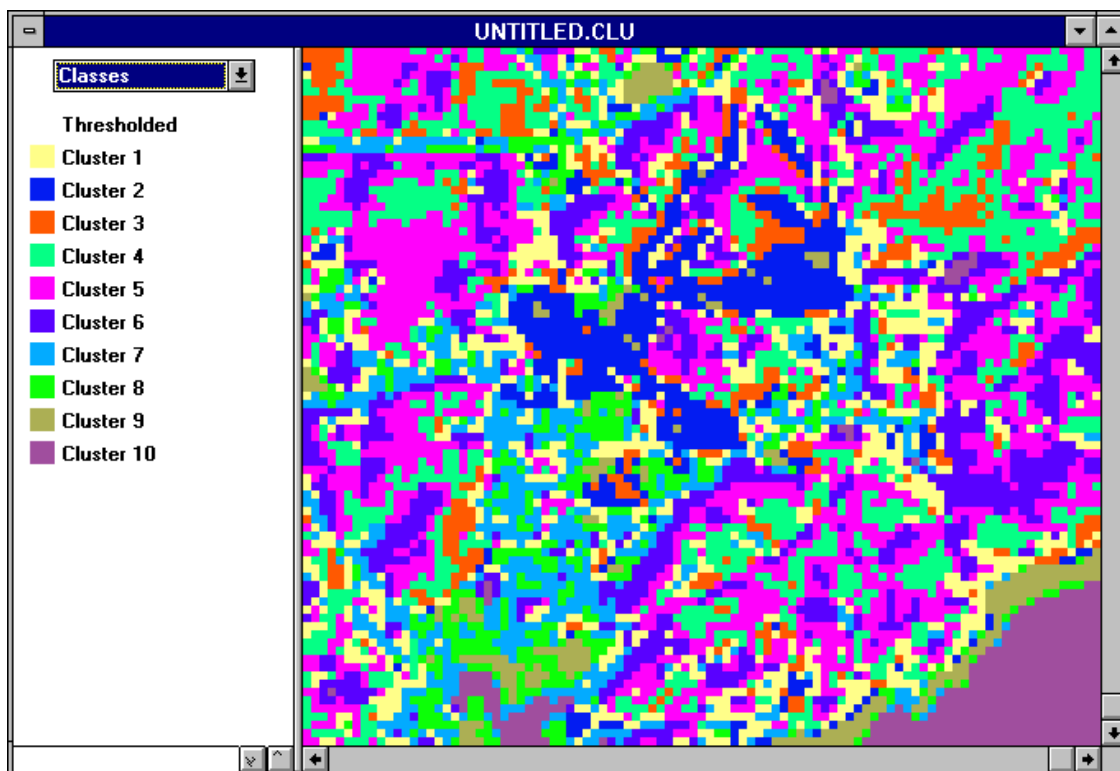
Number classes: 11

Number groups: 0

☒ Display legend

Buttons: Cancel, OK

- Your clustered image opens and, **after Zooming to X5.0**, is shown below.



- Notice that there are 10 numbered classes, plus a class labeled “Thresholded.” This “thresholded” class contains no pixels, because you set the “Thresholding” value to 100, on page 6. Each class is assigned a color by the system which has nothing whatsoever to do with what the cluster represents. The clusters are produced and arranged in order of descending level of brightness. That is, clusters near the top of the list represent surface materials that are “brighter” (have greater reflectance) than those near the bottom of the list.
- The PC version of this software does not yet have the ability (as of the June 6 1997 release) to change the color used to represent a cluster. Future releases may implement this feature.
- You may print the image from the **File** menu. When you do, the clustering key will be printed along with the image.
- You may use some of MultiSpec’s regular tools with this Thematic Map. Such tools as: the **Zoom** feature, and **Coordinate Bar**, from the **View** menu, function normally. The **New Selection Graph** feature will show a plot with only one piece of data. This map is no longer “multispectral.” Each pixel no longer contains data for different Landsat bands, or channels. Each pixel contains only one value, which identifies its color.

- If you do a clustering with a larger number of classes, you may not be able to see them all in the “**Classes**” column. To scroll through this column:
 - Move your cursor into the column
 - Hold the mouse button down
 - Drag to either the top or bottom of the column.

The classes will scroll up and down.

- You and your students will probably want to prepare a thematic map from this clustered image in which you identify some of the clustered areas by their actual land cover. To do this, you may save the image as a **TIFF** file from the **File** menu. This process does not save the clustering key, only the image area will be saved. The TIFF file may then be brought into any one of a number of paint or draw programs to be “fancied up” as a thematic map.
- If you wish to have an image that contains the clustering key, and can also be moved into paint or draw programs you can capture the entire screen using one of a variety of “screen capture” programs that are available in the public domain or as “shareware.” You will want to examine the features of these to determine that they save “captures” in a format that can be read by your paint/draw program.

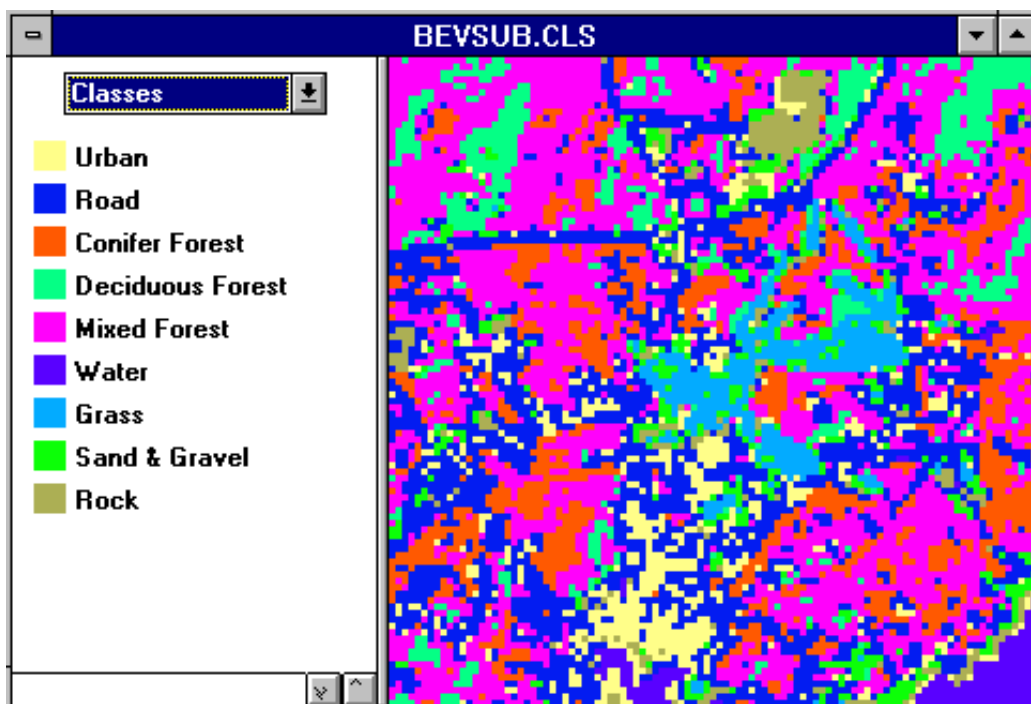
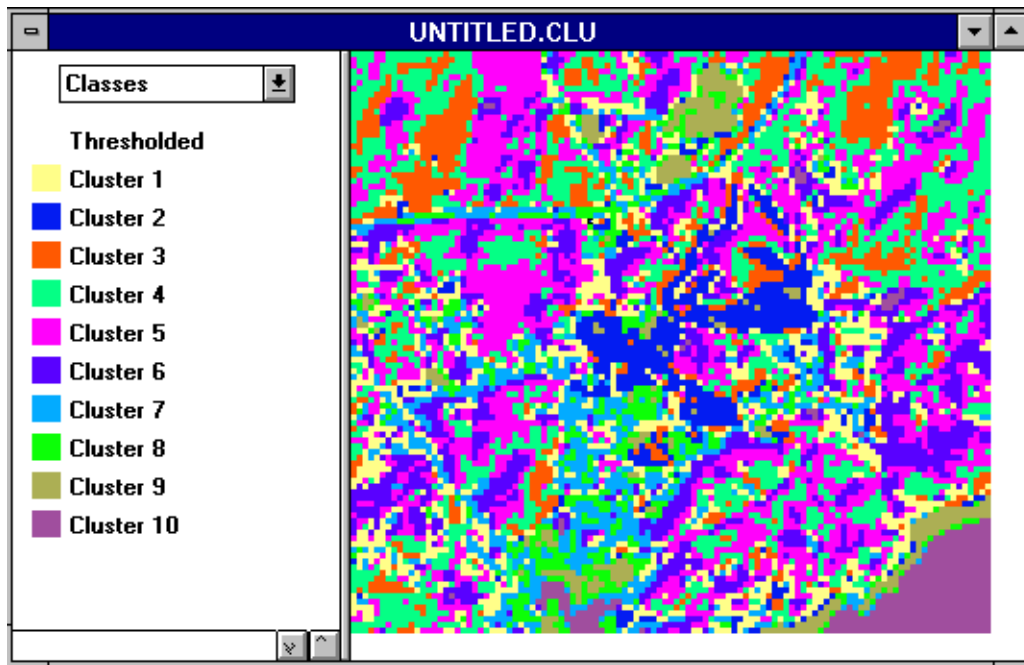
How Valid Is the Clustering Process?

It is necessary for you to be confident that this process of “unsupervised classification” actually yields clusters that are related to land cover types. To this end, included with this tutorial is a file named “**bevsub.cls**”. This is the same image that you have clustered, only this image was prepared with a **supervised** classification by an individual very familiar with the land cover types in the area.

- In your clustered image, zoom in to 3X.
- From the **File** menu, select Open.
- Locate the **bevsub.cls** image, and **Open it**.
- Resize the images and arrange them side-by-side on the screen.

- Compare the areas identified in the supervised classification (the .cls image) to the clusters produced by the system in your unsupervised clustering (the .clu image.)

You should see that the unsupervised clustering provides, at least in this case, a good indication of the locations of large areas of uniform land cover that could be investigated for verification studies, as shown below.



How Many Clusters Do I Use?

Most regions the size of your 15 km x 15 km GLOBE Study Site do not generally demonstrate a large number of different land covers. When you first perform a clustering on your 512 x 512 image, use the same values as you used in this tutorial. Examine the results in light of your knowledge of your own area. Do some field work and look at the areas your clustering suggests are fairly large and homogeneous. Compare your findings to the MUC classification scheme. Only if you feel that this clustering does not adequately represent the land covers in your area should you increase the number of clusters, and then 12 to 14 clusters should be sufficient to do the job.

Reporting the Data

In order to report your data, you must make some “sense” out of the clusters determined by this unsupervised process. You can then re-label the clusters as what type of land cover they represent. The process involves the following steps:

- Desk Verification
- Field Verification
- Completing the Accuracy Assessment for your land cover map*
- Renaming the Clusters
- Sending in your completed map.

(* Before you can submit your map to GLOBE, it is necessary for you and your students to determine how accurately you identified your land cover elements. See the **Accuracy Assessment** section of the Biometry/Land Cover module of your GLOBE Teachers Guide.)

Desk Verification

This process involves your use of local maps (topographic, land cover, soil, political, etc.), other local references (aerial photos, people, agencies, etc.) and the combined experiences of both you and your students to identify some of the clusters produced by MultiSpec. Use whatever resources you can to identify them. Remember that your identifications should correspond to the level IV of the **MUC** (Modified Unesco Classification) scheme.

Field Verification

If there are clusters that you cannot identify “from the desk,” you will have to go out into the field to determine what they are. If a formal “field trip” is not possible, in all probability someone lives near to or drives by that area and can do the identification.

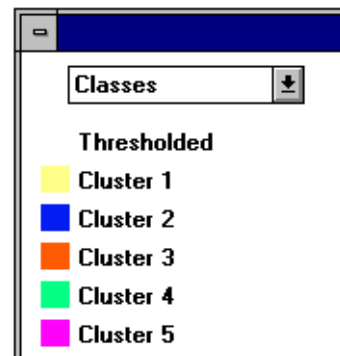
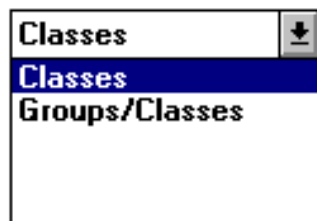
Renaming the Clusters

- Close the .cls image by selecting **Close Window** from the **File Menu**.

Your unsupervised clustering produced clusters identified only by a number, and arranged in order of decreasing brightness. Once you have identified the land cover for each of these clusters, your Thematic Map display may be customized to show these clusters either by name or by **MUC** identification code. You can, in effect, produce two different Thematic Maps on the same image; one in which each cluster is identified by a name (e.g. Ocean, Transportation) and the other by **MUC** designations (e.g. 72, 93.)

The secret to this process is that your **Thematic Map** can display both “**Groups**” and “**Classes**.” When it is produced, both “Groups” and “Classes” have the same set of colors and labels. To see this:

- Click on the **Classes** pull-down menu shown in the diagram to the right.
- The pull-down menu will show the choices illustrated below.
- Select “**Groups/Classes**.” then immediately pull down the window again and select



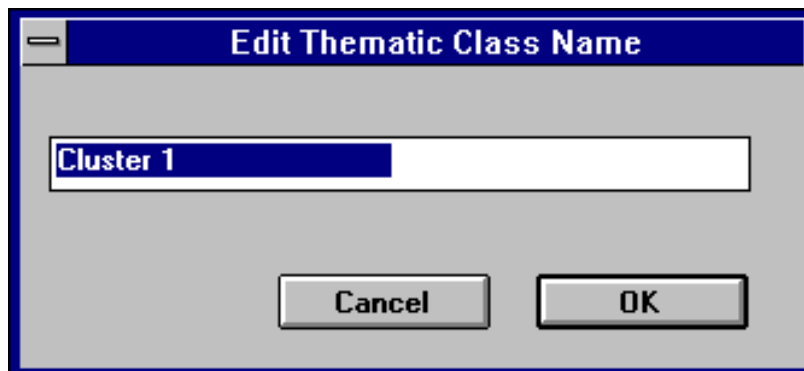
“**Groups**.”

- You can now switch between “**Groups**” and “**Classes**,” you will see that the information in each view is identical.

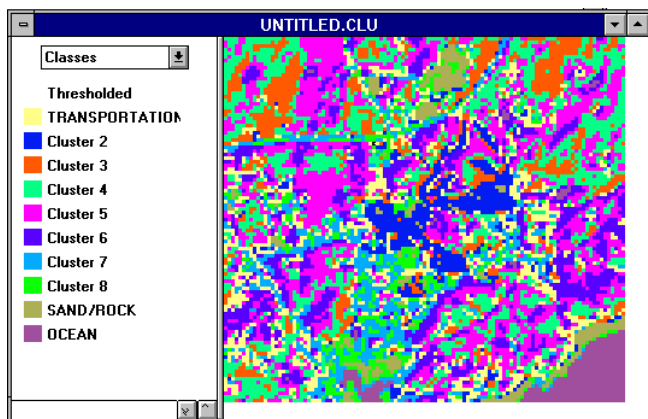
You might decide that the “Groups” will contain descriptive names, while “Classes” contains MUC labes.

To change the name of a cluster in either view, at any time:

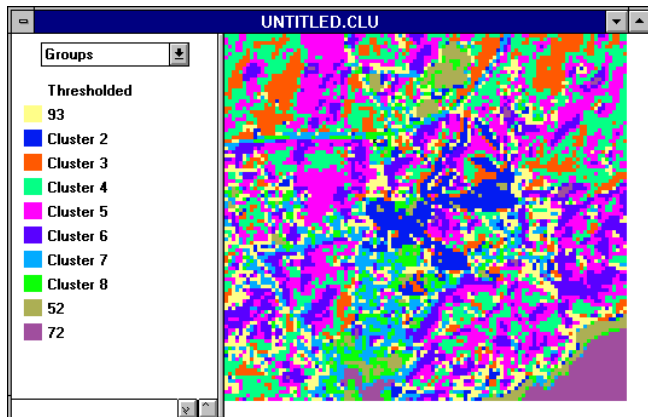
- **Double Click** on the cluster name (for example, “Cluster 1”)
- The **Edit Thematic Class Name** dialog box, as shown below, opens.



- You may now enter either a descriptive name or MUC identification number for this class.
- Once you enter a descriptive name in, say, “Groups,” use the pull-down menu to select “Classes” and enter the corresponding MUC identification number for that same Cluster.



The diagrams to the left show the same image, your bevsb.clu, with some different data displayed in the Classes and Group functions.



You need to be aware that the present version (6 June 1997) of MultiSpec PC only saves data in the **GROUPS**, so be certain that your MUC data is entered here. Once you have entered this data:

- From the **File** menu, select **Save Thematic Group Info**.
- From the **Project** menu, select **Save Project**.
- Accept the default name, and click **OK**.

A Note About Expectations, and a Caveat

When you proceed to the classification of your own 512 x 512 image, you will find the appearance of your clustered image probably considerably different than this demonstration. Major reasons will be:

- a. This sub-set image may not contain as many land cover types as would be found in a full-sized 512 x 512 image.
- b. The nature, abundance and distribution of land cover types in your image will certainly differ from those in the Beverly, Massachusetts area.

As you cluster your own image, you may find that specifying 10 clusters does not discriminate between standing bodies of water, except perhaps between fresh and salt water. In other words, lakes, ponds, rivers, etc. will probably all be clustered into the same group, unless there are significant surface properties that might change their reflectance (i.e., significant algal growth on the surface of a farm pond.)

Validating Your Results

Before submitting your classified land cover map, be certain to perform **Protocol Three: Training and Validation Data Collection** and **Protocol Four: Accuracy Assessment**, found in the **Land Cover/Accuracy Assessment** section of your GLOBE Teacher's Guide.

Submitting your results:

Once you have an unsupervised classification (clustering) that seems to adequately represent your 15 x 15 GLOBE Study Site, your results will be submitted to the **GLOBE Map Archive**, as described in your Teachers Guide.

- Make a copy of your clustered Thematic image onto a high-density floppy disk and clearly label it with your school name, your name, and "clustered image."
- Using your favorite word processor, prepare a file with the following "metadata:"

Your School Name
Your Name
School Address
Date your image was acquired*
The Landsat "path and row" data for your image.*
Some information about yourself, your students, and some of your experiences in doing your clustering.

*This data should be printed on the color prints of your GLOBE Study Site provided to you by the GLOBE Program. From your word processors options, save this data as a text file and place it on the same disk with your

- From your word processors options, save this data as a **text file** (or ASCII file) and place it on the same disk with your Thematic Image.
- Carefully package these disks and send them to:

GLOBE Student Data Archive
NOAA/NGDC E/CG 1
325 Broadway
Boulder, CO, USA 80303